ISRN TELCORDIA--2013-10+PR-0GARAU
Prepared for: Office of Naval Research

Fast Multiscale Algorithms for Information Representation and Fusion

Technical Progress Report No. 10

Devasis Bassu, Principal Investigator

Contract: N00014-10-C-0176

Applied Communication Sciences
150 Mount Airy Road
Basking Ridge, NJ 07920-2021

January 2013

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **JAN 2013**	2. REPORT TYPE	3. DATES COVERED **00-00-2013 to 00-00-2013**

4. TITLE AND SUBTITLE
Fast Multiscale Algorithms for Information Representation and Fusion

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Applied Communication Sciences,150 Mount Airy Road,Basking Ridge,NJ,07920

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR’S ACRONYM(S)

11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In the ninth quarter of the work effort, we focused on a) conducting experiments on real-world data sets using the developed algorithms, b) continued design/implementation of the Multiscale Heat-Kernel Coordinates (MHKC) algorithms and c) packaging for releasing the software as open source. This report documents algorithm designs for the MHKC algorithms. The project is currently on track ? in the upcoming quarter, we will continue applying the developed algorithms to various data sets and the design/implementation of the multiscale heat kernel coordinates algorithms. No problems are currently anticipated.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**12**	

1 Abstract

In the ninth quarter of the work effort, we focused on a) conducting experiments on real-world data sets using the developed algorithms, b) continued design/implementation of the Multiscale Heat-Kernel Coordinates (MHKC) algorithms and c) packaging for releasing the software as open source. This report documents algorithm designs for the MHKC algorithms.

The project is currently on track – in the upcoming quarter, we will continue applying the developed algorithms to various data sets and the design/implementation of the multiscale heat kernel coordinates algorithms. No problems are currently anticipated.

Table of Contents

2 Summary

In this quarter, we continued design and implementation of the new multiscale heat kernel coordinates (MHKC) algorithms. The current design variants for MHKC algorithms are documented in this report.

The project is currently on track – in the upcoming quarters, we will continue applying the developed algorithms to various data sets and focus on the design and development of the MHKC algorithms. No problems are currently anticipated.

3 Introduction

The primary project effort over the last quarter focused on completing design/development of the multiscale heat-kernel coordinates algorithms [1]. This provides a power tool for discovering the non-linear geometries in any given dataset. This utilizes the fast randomized Singular Value Decomposition (RSVD) algorithms described in the earlier ONR reports [7][8]. Use of the RSVD effectively reduces the computational complexity from $O(\text{m.n.k})$ to $O((\text{m+n}).\text{k}^2)$ for an *m* by *n* matrix of rank *k*. In contrast to the multiscale Singular Value Decomposition (MSVD) algorithms that detect linear structures in data at multiple scales, the MHKC uses heat kernels to discover the non-linear manifold structure in which the data resides at various scales. Similar to the MSVD, the MHKC provides an efficient representation using low-dimensional coordinates corresponding to the original data points.

An outline of the MHKC algorithm was presented in the previous quarterly report [10]. While most of the algorithm is automated, the crucial step of selecting the *appropriate* heat-kernel coordinates for any given application required manual intervention on part of the data analyst. In this report, we present two canonical approaches to automating the selection of the MHKC embedding. Further, it also provides a way to visualize the embedded data in lower (2 or 3) dinmensions.

4 Methods, Assumptions and Procedures

4.1 Multiscale Heat Kernel Coordinates

The Multiscale Heat Kernel Coordinates (MHKC) algorithms are based on theoretical results presented in [1]. The current algorithm design is described below.

Input: A set of *n* data points $\{x_1, x_2, \dots, x_n\}$ in R^d. Assume *n* is large.

Step 1 (Normalization): Normalize the points x_i such that the data cloud is in a ball of unit variance. Define

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|x - \vec{x}\|^2}$$

where $\vec{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. The normalized point y_i corresponding to x_i is given by

$$y_i = \frac{x_i - \vec{x}}{\sigma_x}$$

Note: The translation to mean zero is not necessary for the purposes of building the transition probability matrix in the next step.

Step 2 (Transition Probability Matrix): The second step comprises constructing the data matrix to be provided as input to the RSVD algorithm. Define the heat kernel as

$$k(x,y) = exp\left(-\frac{\|x-y\|_2^2}{t_0}\right)$$

for any two points *x* and *y*. Here, t_0 is a constant (data dependent) representing the kernel window size (set $t_0 = 2^{-s}{\sigma_y}^2$ for scale $s \geq 0$; select *s* representing some finer scale of interest). The heat kernel matrix is then defined as

$$K=\{k_{ij}\} \text{ where } k_{ij} = k(x_i, x_j)$$

for $i, j = 1,2,\dots,n$. The transition probability matrix is $P = D^{-1}K$ where *D* is the diagonal matrix with the *i*-th entry as sum of the *i*-th row of *K*.

Note: For large n, compute $\beta \approx 25$ elements for each row of K using the randomized approximate nearest neighbor algorithm ([9]). This reduces the computational complexity from $O(n^2.d)$ to $O(n.log(n).d)$ and captures local information.

Note that P is not symmetric. There are various techniques to symmetrize P such that the eigenvalues and eigenfunctions are still easy to compute. One way is to define

$$P' = D^{-1/2}.P.D^{-1/2}$$

P' is symmetric with the same eigenvalues as P. Also, the eigenvectors can easily be easily obtained using a simple transformation of either $D^{-1/2}$ or $D^{1/2}$. The RSVD algorithm may be used to compute the spectrum of P'.

Step 3 (MHKC Embedding): Next, the heat kernel coordinates is defined for each of the original data points. Let the eigenvalues of P be defined as λ_j and the right-eigenvectors as v_j for j = 1,2,…,rank(P).

Each point x_i is then represented as $HKC(x_i) = (exp(-\lambda_1 t).v_{11}, exp(-\lambda_2 t).v_{21}, \ldots, exp(-\lambda_r t).v_{r1})$ where v_{ji} is the i-th coordinate of the eigenvector v_j. Here t is the time/scale parameter that is to be varied to look at the geometries of the data set at various scales.

Note: The first eigenvalue/eigenvector of P is trivial and should not be used.

Next, we provide two approaches to automating the selection of heat-kernel coordinates in Step 2 of the algorithm described above.

4.1.1 Representation Using Canonical Clustering

Choose a small integer l representing the number of non-trivial eigenvectors v_j to consider (obtained via the algorithm in Section 4.1). Construct a binary tree using the first l non-trivial eigenvectors to divide the n points into $N_c = 2^l$ clusters as follows. First, divide the points into two sets defined by

$$\left\{x_i \mid y_i^1 \geq \frac{1}{n}\sum_{j=1}^{n} y_j^1\right\} \text{ and } \left\{x_i \mid y_i^1 < \frac{1}{n}\sum_{j=1}^{n} y_j^1\right\}$$

where $y_i^k = v_k(x_i)$ using the first diffusion vector. Apply the 2nd diffusion vector to both these sets; repeat recursively to obtain N_c disjoint clusters. For each cluster $c = 1, 2, \ldots, N_c$, find the closest point $v\left(x_{\mu(c)}\right)$ to the cluster mean.

Select a suitable time t (same for all the clusters) and use the heat kernel $k(x, y)$ to get a vector of length N_c given by

$$z_i = \left(k\left(x_i, x_{\mu(1)}\right), k\left(x_i, x_{\mu(2)}\right), \ldots, k\left(x_i, x_{\mu(N_c)}\right)\right)$$

for each x_i. Now, perform standard PCA on the dataset $\{z_i\}$. To visualize the dataset, simply use projections on the first 2 or 3 dimensions.

4.1.2 Representation Using Multiscale SVD

For each k, normalize $(y_1^k, y_2^k, \ldots, y_n^k)$ to have unit length. Define $z_i^k = y_i^k / \sum_{j=1}^{n} y_j^k$. Pick a small integer, say 2. Out of the top ten non-trivial diffusion eigenvectors, choose the two values k_1, k_2 such that the set (in 2D) consisting of all the points

$$\{(z_i^{k_1}, z_i^{k_2}) \mid i = 1,2,\ldots,n\}$$

has the smallest average value of multiscale SVD at scales 2^{-s} for $s = 0,1,2,3$ (at each location and scale, compute the average squared distance to the best fitting line). Amongst the various "best" choices, pick the one with the smallest value of $k_1 + k_2$.

4.2 Deliverables / Milestones

Date	Deliverables / Milestones	Status
Oct 2010	Progress report for period 1, 1st quarter	✓
Jan 2011	Progress report for period 1, 2nd quarter / complete randomized matrix decompositions task	✓
Apr 2011	Progress report for period 1, 3rd quarter / complete approximate nearest neighbors task	✓
Jul 2011	Progress report for period 1, 4th quarter / complete experiments – part 1	✓
Oct 2011	Progress report for period 2, 1st quarter	✓
Jan 2012	Progress report for period 2, 2nd quarter / complete multiscale SVD task	✓
Apr 2012	Progress report for period 2, 3rd quarter	✓
Jul 2012	Progress report for period 2, 4th quarter / complete experiments – part 2	✓
Oct 2012	Progress report for period 3, 1st quarter	✓
Jan 2013	Progress report for period 3, 2nd quarter / complete multiscale Heat Kernel task	✓
Apr 2013	Progress report for period 3, 3rd quarter	
Jul 2013	Final project report + software + documentation on CDROM / complete experiments – part 3	

5 Results and Discussion

We described two approaches to automate the process of selecting the “appropriate” diffusion vectors for a given dataset. The first approach in itself provides an agnostic and canonical way of “clustering”. In terms of computational cost, the first approach is much better as it avoids the potential combinatorial explosion in the second approach. However, the second approach directly evaluates the information content for each diffusion vector in a multiscale sense and picks the “best” combination. We will experimentally evaluate both these techniques against real-world datasets.

6 Conclusions

The project is on track with design/implementation of the new multiscale heat kernel coordinates algorithms. We will continue with algorithmic improvements and experimentation using the developed algorithms in the next quarter.

No problems are currently anticipated.

7 References

[1] P.W. Jones, M. Maggioni, R. Schul, *Manifold parametrizations by eigenfucntions of the Laplacian and heat kernels*, PNAS, vol. 105, no. 6, pp. 1803-1808.

[2] E. Liberty, F. Woolfe, P.G. Martinsson, V. Rokhlin, M. Tygert, *Randomized Algorithms for the Low-Rank Approximation of Matrices*, PNAS, vol. 104, no. 51, pp. 20167-20172, 2007.

[3] E. Liberty, F. Woolfe, P.G. Martinsson, V. Rokhlin, M. Tygert, *A Fast Randomized Algorithm for the Approximation of Matrices*, Applied and Computational Harmonic Analysis, vol. 25, pp.335-366, 2008.

[4] V. Rokhlin, M. Tygert, A fast Randomized Algorithm for the Overdetermined Linear Least Squares Regression, PNAS, vol. 105, no. 36, pp. 13212-13217, 2008.

[5] G. H. Golub, W. Kahan, *Calculating the singular values and pseudo-inverse of a matrix*, Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis, vol. **2** (2), pages 205–224, 1965.

[6] G.H. Golub, C.F. Van Loan, *Matrix Computations (3rd edition.),* Johns Hopkins University Press, Baltimore, 1996.

[7] D.Bassu, *Fast Multiscale Algorithms for Information Representation and Fusion, Technical Report No. 1*, ISRN TELCORDIA--2010-01+PR-0GARAU, 2010.

[8] D.Bassu, *Fast Multiscale Algorithms for Information Representation and Fusion, Technical Report No. 2*, ISRN TELCORDIA—2011-02+PR-0GARAU, 2011.

[9] D.Bassu, *Fast Multiscale Algorithms for Information Representation and Fusion, Technical Report No. 3*, ISRN TELCORDIA—2011-03+PR-0GARAU, 2011.

[10] D.Bassu, *Fast Multiscale Algorithms for Information Representation and Fusion, Technical Report No. 9*, ISRN TELCORDIA—2012-09+PR-0GARAU, 2012.